To Weight or not to Weight: Source-Normalised LDA for Speaker Recognition using i-vectors

Mitchell McLaren and David van Leeuwen

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

{m.mclaren, d.vanleeuwen}@let.ru.nl

Abstract

Source-normalised Linear Discriminant Analysis (SN-LDA) was recently introduced to improve speaker recognition using i-vectors extracted from multiple speech sources. SN-LDA normalises for the effect of speech source in the calculation of the between-speaker covariance matrix. Source-normalised-and-weighted (SNAW) LDA computes a weighted average of source-normalised covariance matrices to better exploit available information.

This paper investigates the statistical significance of performance gains offered by SNAW-LDA over SN-LDA. An exhaustive search for optimal scatter weights was conducted to determine the potential benefit of SNAW-LDA. When evaluated on both NIST 2008 and 2010 SRE datasets, scatter-weighting in SNAW-LDA tended to overfit the LDA transform to the evaluation dataset while offering few statistically significant performance improvements over SN-LDA.

Index Terms: speaker recognition, linear discriminant analysis, i-vector, source variability

1. Introduction

Source-normalised (SN) Linear Discriminant Analysis (LDA) [1] provides significant improvements over the standard approach to LDA when dealing with mismatched trial conditions in speaker recognition based on i-vectors [2]. This technique was developed to assist in speaker recognition when dealing with multiple speech sources (such as telephone, microphone and interview sources) and unseen combinations of these sources in trials. SN-LDA accomplishes this objective by firstly estimating between-speaker variation on a source-conditioned basis prior to accumulating the final, source-normalised between-speaker scatter matrix. The within-speaker variation is then given by the residual variability in the i-vector space not observed as between-speaker variation.

It was shown in [1] that weighting the source-conditioned between-class scatters toward the most reliably estimated scatter provided additional performance improvements over SN-LDA. This approach, termed source-normalised-and-weighted (SNAW) LDA, calculates a weighted average of the scatters in an heuristic manner based on the proportion of the LDA training dataset used in the estimation of each source-conditioned scatter. While this weighting proved effective in [1], it was trialled on a single dataset composition which was heavily biased toward telephone-sourced speech and using a sub-optimal system configuration, namely, a 512-component universal background model (UBM) and standard MFCC extraction parameters.

This research was funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 238803.

This paper extends on the initial development of the source-normalised LDA algorithms by analysing the significance of performance improvements offered by SNAW-LDA over SN-LDA using an improved i-vector configuration and multiple LDA training dataset compositions. The potential benefit of SNAW-LDA is evaluated using an exhaustive search of *oracle* scatter weighting parameters to maximise system performance on the recent NIST 2008 and 2010 speaker recognition evaluation (SRE) corpora. The oracle weights are then analysed to provide some insight into the ability of SNAW-LDA transforms to generalise to unseen data.

This paper is structured as follows. Section 2 briefly describes the i-vector framework for speaker recognition. Section 3 contrasts the SN- and SNAW-LDA algorithms against the standard LDA approach. The experimental protocol and corresponding results are given in Sections 4 and Section 5.

2. The i-vector Speaker Recognition Framework

The i-vector framework [3] consists of three stages — i-vector extraction, inter-session compensation and classification using a cosine kernel function.

I-vector Extraction: An i-vector is a compact representation of a speech utterance extracted from a space in which the majority of between-utterance variability is expected to lie. This space $\boldsymbol{T}$ is referred to as the *total variability subspace*. The total variability subspace assumes that an utterance can be represented by the Gaussian mixture model (GMM) mean supervector, $\boldsymbol{M}$=$\boldsymbol{m}$+$\boldsymbol{T}\boldsymbol{w}$, where $\boldsymbol{M}$ consists of a speaker- and session-independent mean supervector $\boldsymbol{m}$ from the universal background model (UBM) and a mean offset $\boldsymbol{T}\boldsymbol{w}$. The low-rank i-vector $\boldsymbol{w}$ has a standard normal distribution $\mathcal{N}(0,1)$ and is found through maximum a-posteriori adaptation of $\boldsymbol{M}$ in the space defined by $\boldsymbol{T}$. Details on the constraints of model parameters and an efficient algorithm for estimating $\boldsymbol{T}$ and $\boldsymbol{w}$ can be found in [3].

Inter-session Compensation: Due to the nature of the total variability subspace, i-vectors exhibit both speaker-intrinsic or speaker-extrinsic variation. I-vectors are, therefore, subject to inter-session variability compensation prior to classification in order to optimise speaker discrimination. Both LDA and WCCN are utilised for this purpose in the i-vector framework. LDA aims to find a reduced set of axes $\boldsymbol{A}$ that maximises the between-speaker variability observed in the i-vector space while simultaneously minimising the within-speaker variability. The LDA process is covered in detail in Section 3. The secondary stage, Within-Class Covariance Normalisation (WCCN) [4], normalises the residual within-speaker variance

remaining in LDA-reduced i-vectors. The WCCN matrix $\boldsymbol{B}$ is found through the Cholesky decomposition of the within-class covariance as described by [4].

Cosine Distance Scoring: A trial score is given by the cosine distance between a set of i-vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. The cosine distance is is the dot product $\langle \hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2 \rangle$ of the inter-session-compensated and normalised vectors,

$$\hat{\boldsymbol{w}}_i = \frac{\boldsymbol{B}^t \boldsymbol{A}^t \boldsymbol{w}_i}{\|\boldsymbol{B}^t \boldsymbol{A}^t \boldsymbol{w}_i\|}. \quad (1)$$

In this work, cosine kernel normalisation [5] is additionally employed.

3. Source-Normalisation in LDA

LDA is an important component in the i-vector framework for speaker recognition. LDA serves the purpose of enhancing discrimination between i-vectors corresponding to different speakers. This section describes the standard LDA algorithm and the recently proposed SN- and SNAW-LDA variants. Each of these algorithms attempt to minimise the within-speaker variability observed in a training dataset while maximising the between-speaker variability through the generalised eigenvalue decomposition of $\boldsymbol{S}_B \boldsymbol{v} = \lambda \boldsymbol{S}_W \boldsymbol{v}$ where $\boldsymbol{S}_B$ represents the between-speaker scatter and $\boldsymbol{S}_W$ the within-speaker covariance matrix.

3.1. Linear Discriminant Analysis

In the standard approach to LDA, the scatter matrices $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ are calculated as,

$$\boldsymbol{S}_B = \sum_{s=1}^{S} N_s (\boldsymbol{\mu}_s - \boldsymbol{\mu})(\boldsymbol{\mu}_s - \boldsymbol{\mu})^t \quad (2)$$

$$\boldsymbol{S}_W = \sum_{s=1}^{S} \sum_{i=1}^{N_s} (\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)(\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)^t. \quad (3)$$

The number of i-vectors from speaker s is given by N_s and $\boldsymbol{\mu}_s$ represents the mean of these i-vectors. The i-vector mean $\boldsymbol{\mu} = 0$ due to the assumption of normally distributed and zero-mean i-vectors [3, 6]. A recent study [1] has highlighted that eq. (2) and eq. (3) are often poorly estimated in the context of speaker recognition. This is because each speaker in a typical training dataset does not provide speech from each source of interest. In this scenario, the within-speaker variation is not estimated in its entirety which subsequently allows residual variation to adversely influence the estimation of the between-speaker scatter. SN-LDA was proposed in the same study to counteract these shortcomings.

3.2. Source-Normalised LDA

The influence of source-related variation on the between-speaker scatter can be reduced by estimating this scatter using source-normalised vectors that are calculated with respect to their corresponding source mean. The source-normalised $\overline{\boldsymbol{S}_B}$ can then be composed of the source-normalised scatter matrices such that

$$\overline{\boldsymbol{S}_B} = \sum \boldsymbol{S}_B^{\mathrm{src}} \quad (4)$$

$$\boldsymbol{S}_B^{\mathrm{src}} = \sum_{s=1}^{S_{\mathrm{src}}} N_{\mathrm{src}} (\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\mathrm{src}})(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\mathrm{src}})^t, \quad (5)$$

where $\boldsymbol{\mu}_{\mathrm{src}} = \frac{1}{N_{\mathrm{src}}} \sum_{n=1}^{N_{\mathrm{src}}} \boldsymbol{w}_n^{\mathrm{src}}$ and N_{src} designates the number of speech samples taken from source src.

Since $\overline{\boldsymbol{S}_B}$ does not bound variation due to the different speech sources, this variation, along with all other factors contributing within-speaker variability in the training dataset, can be estimated as the residual variation in the i-vector space. Given that $\boldsymbol{S}_T = \sum_{n=1}^{N} (\boldsymbol{w_n} - \boldsymbol{\mu})(\boldsymbol{w_n} - \boldsymbol{\mu})^t$ represents the total variance observed in the i-vector space (note that the source-independent i-vector mean $\boldsymbol{\mu}$ is a null vector), and $\overline{\boldsymbol{S}_B}$ the source-normalised estimate of the between-speaker scatter, the within-speaker covariance matrix is formulated as

$$\boldsymbol{S}_W = \boldsymbol{S}_T - \overline{\boldsymbol{S}_B}, \quad (6)$$

In taking this approach, the accurate estimation of the within-speaker scatter is no longer dependent on the availability of multi-source utterances per speaker as is the case with eq. (3). Readers are directed to [1] for more information on the SN-LDA algorithm.

3.3. Source-Normalised-And-Weighted LDA

Equation (4) was extended in [1] to incorporate a heuristic-based weighting scheme to bias the final between-speaker scatter $\overline{\boldsymbol{S}_B}$ toward the most reliably estimated source-normalised covariance matrix $\boldsymbol{S}_B^{\mathrm{src}}$. The most reliable representation of between-speaker scatter was deemed to be calculated from the largest collection of i-vectors. The source-normalised-and-weighted (SNAW) between-speaker scatter matrix is thus calculated as,

$$\overline{\boldsymbol{S}_B} = \sum \alpha_{\mathrm{src}} \boldsymbol{S}_B^{\mathrm{src}}. \quad (7)$$

where the weight $\alpha_{\mathrm{src}} = \frac{N_{\mathrm{src}}}{N}$ reflects the total proportion of i-vectors available from source src in the LDA training dataset.

4. Experimental Protocol

Experiments were performed using the recent NIST 2008 and 2010 SRE corpora. Results are reported for four evaluation conditions on each corpora thus following the protocol of [1]. The SRE'10 conditions correspond to det conditions 2–5 in the evaluation plan, and in the case of SRE'08, det conditions 3–5 and 7 [7]. These conditions include a combination of telephone, microphone and interview-sourced English speech. The *extended* evaluation protocol was used for SRE'10 trials in which the number of trials exceed 2.8 million in the case of the *int-int* condition thus providing an adequate number of scores to analyse the statistical significance of results. Performance was evaluated using the equal error rate (EER) and a normalised minimum decision cost function (DCF) calculated using $C_{\mathrm{M}} = 1$, $C_{\mathrm{FA}} = 1$ and $P_{\mathrm{T}} = 0.001$ for SRE'10 results and $C_{\mathrm{M}} = 10$ and $P_{\mathrm{T}} = 0.01$ for SRE'08 results. When contrasting performance statistics, they are deemed significantly different according to Eq.(13) in [8] using a 95% confidence interval around the EER (and an equivalent formula for minimum DCF). In all approaches, the number of LDA dimensions retained was evaluated in steps of 50 in order to minimise the average system errors using $(\mathrm{DCF} + 10 \times \mathrm{EER})$ across the evaluated conditions of SRE'10 and $(\mathrm{DCF} + \mathrm{EER})$ for SRE'08. On average, 200 LDA dimensions were retained.

Gender-dependent, 2048-component UBMs were trained using 60-dimensional, feature-warped MFCCs (including deltas and double-deltas) extracted from the NIST 2004, 2005, and 2006 SRE corpora and LDC releases of Fisher English, Switchboard II: phase 3 and Switchboard Cellular (parts 1 and 2).

Corpus	LDA Algorithm	tel-int		int-tel		int-mic		int-int		tel-tel	
		DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER
SRE'08	Standard	.0231	4.70%	.0243	4.56%	—	—	.0138	3.35%	.0157	3.18%
	SN	.0129	3.27%	.0151	2.91%	—	—	.0117	3.17%	**.0139**	**2.61%**
	SNAW	.0132	3.07%	.0144	**2.55%**	—	—	.0113	3.08%	.0147	2.77%
	SNAW (oracle)	**.0124**	**3.00%**	**.0142**	2.66%	—	—	**.0109**	**2.97%**	.0141	2.77%
SRE'10	Standard	—	—	.6007	4.19%	.4314	2.65%	.5632	3.99%	.5239	3.26%
	SN	—	—	.4559	**2.93%**	.3561	2.09%	.5069	3.14%	.5137	**3.14%**
	SNAW	—	—	**.4497**	2.96%	.3601	**1.81%**	.5075	**2.92%**	.5327	3.17%
	SNAW (oracle)	—	—	.4771	2.95%	**.3275**	1.85%	**.4659**	3.00%	**.4889**	3.21%

Table 1: SRE'10 (extended protocol) results using standard, source-normalised and SNAW LDA approaches.

Voice activity detection was implemented as in [1]. Gender-conditioned datasets were used as data for total variability subspace, LDA and WCCN training and for cosine kernel normalisation. These consisted of the aforementioned corpora along with additional interview data taken from the NIST 2008 SRE follow-up corpus for use in SRE'10 evaluations and from a subset of the 3-minute interview segments of the NIST 2010 SRE corpus for SRE'08 evaluations. The total segment counts N_{tel}, N_{mic} and N_{int} for the gender-conditioned datasets were 30355, 5330, and 2537, respectively.

5. Results

The following experiments investigate the potential benefits of incorporating weighted between-speaker scatter matrices in the source-normalised LDA algorithm via SNAW-LDA. The significance of performance improvements due to SNAW-LDA is analysed in the context of heuristic-based and oracle scatter weightings. Finally, the effect of varying dataset compositions on the oracle weights is analysed to determine the potential for the SNAW-LDA transform to generalise to unseen data.

5.1. Source-Normalised LDA: To weight or not to weight?

This section provides a comparison of standard LDA, SN-LDA and SNAW-LDA with particular focus on the effect of the heuristic-based weighting scheme in SNAW-LDA. Results on the SRE'08 and SRE'10 corpora when using these algorithms are provided in Table 1. SN-LDA offered relative improvements of 31–44% over standard LDA in the mis-matched trials of SRE'08. Similarly, SN-LDA consistently provided statistically significant improvements over standard LDA in the SRE'10 trials with 10–30% relative improvements observed in trials involving microphone or interview speech. These trends reflect the findings of previous studies [1, 9]. In contrast to these studies, however, SNAW-LDA provided only comparable performance to SN-LDA with statistically significant improvements being observed only in several SRE'10 results. It is believed that SNAW-LDA perhaps counteracted an underlying problem in the sub-optimal system used in previous work and that this shortcoming was no longer exhibited in the improved system configuration in this work — namely, additional UBM components and the use of a wider bandpass filter and more spectral bands during MFCC extraction.

The above trends suggest that with a suitable system configuration, the weighting of between-speaker scatter matrices via SNAW-LDA are of little benefit as the majority of performance improvements are delivered through source-normalisation alone. One explanation for this observation is that the current heuristic-based weighting algorithm for SNAW-LDA is sub-optimal. The following section aims to determine whether an improved weighting scheme is possible through an exhaustive search of a set of oracle scatter weights.

5.2. The Potential of Weighted Scatter in SNAW-LDA

The current SNAW-LDA scatter weighting algorithm is based on heuristics such that the the weight α associated with scatter $\boldsymbol{S}_B^{\text{src}}$ is calculated as the proportion of the total training data that originates from source src (see eq. (7)). As identified in the previous section, this weighting algorithm provides no significant benefit to the use of unweighted SN-LDA. To determine whether the development of a new weighting algorithm is worthy of pursuit, oracle weights $\{\alpha_{\text{tel}}, \alpha_{\text{mic}}, \alpha_{\text{int}}\}$ were found to minimise average system errors (see Section 4) on each corpora using an exhaustive search with a resolution 0.1 and under the constraint $\sum_{\text{src}} \alpha_{\text{src}} = 1$. Results when using the oracle weights are provided in Table 1.

An average relative improvement (across conditions) of 4% was observed due to SNAW-LDA (oracle) over SN-LDA in both SRE'08 and SRE'10. SNAW-LDA (oracle) provided no statistically significant improvement over SN-LDA in the SRE'08 results and in most of the SRE'10 trials involving telephone-sourced i-vectors. Given that nearly 80% of the LDA training dataset is sourced from telephone speech, this latter observation brings into question how the composition of the LDA training dataset influences the potential performance offered through scatter weighting in conjunction with SN-LDA. To answer this question, the influence of telephone speech on the LDA training dataset was reduced by scaling the total number of available speakers of telephone utterances by factors $\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ while keeping the number of i-vectors from alternate speech sources constant. Subsequently, the smallest proportion of telephone-sourced i-vectors in the reduced LDA training dataset fell to as low as 10%. Figure 1 illustrates the average SRE'10 minimum DCF across trial conditions for each of the LDA training algorithms with respect to the proportion of available telephone-sourced i-vectors in the training dataset.

Figure 1 shows that for all LDA algorithms (with the exception of SNAW-LDA), the minimum DCF was largely unaffected by the reduction of telephone-sourced training i-vectors. A significant and consistent performance improvement was also observed from SN-LDA over the standard approach to LDA. These observations extended across all performance statistics in both corpora as the size of the training dataset was reduced. Figure 1 further indicates that SNAW-LDA typically provided worse performance than SN-LDA while SNAW-LDA (oracle) offered consistent improvements over SN-LDA. Although not

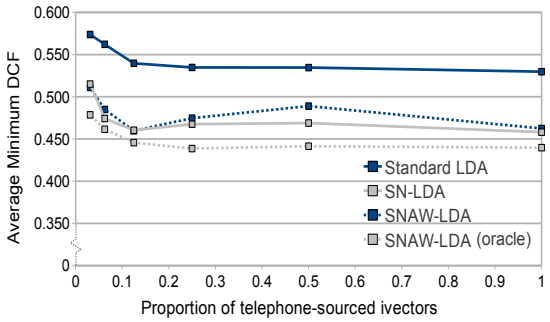

Figure 1: Average SRE'10 min. DCF for each LDA variant as telephone i-vectors were reduced from the training dataset.

shown here, the corresponding plots for SRE'10 EER and SRE'08 DCF and EER indicated, however, that comparable performance was offered from each of the SN-LDA variants. The reason for the anomalies observed in the SRE'10 minimum DCF when using different SN-LDA variants is further analysed with respect to weight allocation in the following section.

5.3. Weight Analysis

This section analyses the distribution of heuristic-based and oracle SNAW-LDA weights for SRE'08 and SRE'10 trials with the objective of determining whether weights tend to be specific to the evaluation dataset thus providing an explanation for the inconsistencies observed across copora in the previous section. Figure 2 depicts these distributions for the (a) heuristic, (b) oracle SRE'08, and (c) oracle SRE'10 weights allocated to each source as the amount of telephone-sourced observations in the LDA training dataset was reduced from 100% at the top of the plot to a proportion of $\frac{1}{32}$ at the bottom.

Several interesting trends can be observed from Figure 2. Firstly, the SRE'08 oracle weights followed a broadly similar trend to heuristic-based weights in that the interview scatter weight increased as telephone i-vectors were removed from the dataset. In contrast, the SRE'10 evaluations found benefit from a more static allocation of weights with preference towards the interview scatter. Secondly, and perhaps more importantly, is the observation that some scatters were allocated an oracle weight of zero. This was particularly evident in the SRE'08 oracle weights. While a weight of zero may have provided best overall performance on the given evaluation corpus, this has the effect of severely limiting the ability for the resulting LDA transform to generalise to unseen data. Specifically, SNAW-LDA calculates the within-speaker scatter as the residual variation in the i-vector space after the estimation of the weighted between-speaker scatter (see eq. (6)). Consequently, a scatter weight of zero implicitly causes the SNAW-LDA algorithm to treat the corresponding between-speaker information as within-speaker variation, thus the between-speaker information that has potential to improve speaker discrimination in alternate copora is suppressed rather than exploited in the LDA optimisation. These findings suggest, therefore, that the weighting scatters in SNAW-LDA provides a means of tuning the LDA transform to the conditions of the evaluation corpus.

6. Conclusion

This study analysed the performance offered by the SNAW-LDA algorithm compared to more straightforward SN-LDA.

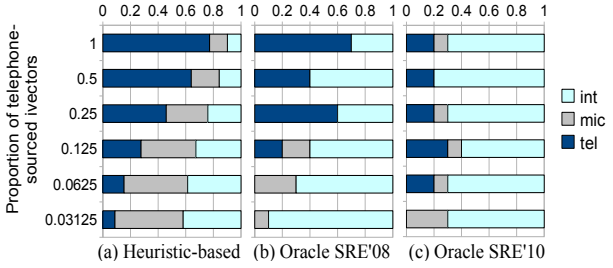

Figure 2: Distribution of weights allocated to the SN between-speaker scatters in SNAW-LDA as the number of telephone i-vectors were reduced from the LDA training dataset.

Particular focus was placed on the limited statistical significance of improvements offered by SNAW-LDA over SN-LDA on both the SRE'08 and SRE'10 evaluation corpora. It was demonstrated that SN-LDA provided significant and consistent improvements over the use of standard LDA across different LDA training dataset compositions. Despite finding oracle sets of scatter weights for SNAW-LDA, improvements over SN-LDA remained limited. Further, weights of zero were allocated to several between-speaker scatters in the oracle experiments thereby hindering the ability of the resulting LDA transform to generalise to unseen data. These trends suggest that the weighting of scatter matrices is essentially tuning the LDA transform to match the evaluation conditions. Consequently, it would seem that the straightforward and robust SN-LDA is a more appropriate solution to inter-session compensation in the context of LDA training datasets consisting of speech from multiple sources.

7. References

[1] M. McLaren and D. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 5456–5459.

[2] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. Interspeech*, 2009, pp. 1559–1562.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *In print IEEE Trans. Audio, Speech and Language Processing*, 2010.

[4] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Ninth Int. Conf. on Spoken Language Processing*, 2006, pp. 1471–1474.

[5] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. Odyssey Workshop*, 2010.

[6] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey Workshop*, 2010, pp. 28–33.

[7] National Institute of Standards and Technology, *NIST Speaker Recognition Evaluation website*, available: http://www.itl.nist.gov/iad/mig/tests/sre/.

[8] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. Odyssey Workshop*, 2004.

[9] M. McLaren and D. van Leeuwen, "Improved speaker recognition when using i-vectors from multiple speech sources," in *In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 5460–5463.